Rethinking Feature Extraction: Gradient-based Localized Feature Extraction for End-to-End Surgical Downstream Tasks

Winnie Pang[1†], Mobarakol Islam[2†], Sai Mitheran[3], Lalithkumar Seenivasan[1], Mengya Xu[1], Hongliang Ren[1,4*]
Senior Member, IEEE

***Abstract*—Several approaches have been introduced to understand surgical scenes through downstream tasks like captioning and surgical scene graph generation. However, most of them heavily rely on an independent object detector and region-based feature extractor. Encompassing computationally expensive detection and feature extraction models, these multi-stage methods suffer from slow inference speed, making them less suitable for real-time surgical applications. The performance of the downstream tasks also degrades from inheriting errors of the earlier modules of the pipeline. This work develops a detector-free gradient-based localized feature extraction approach that enables end-to-end model training for downstream surgical tasks such as report generation and tool-tissue interaction graph prediction. We eliminate the need for object detection or region proposal and feature extraction networks by extracting the features of interest from the discriminative regions in the feature map of the classification models. Here, the discriminative regions are localized using gradient-based localization techniques (e.g. Grad-CAM). We show that our proposed approaches enable the real-time deployment of end-to-end models for surgical downstream tasks. We extensively validate our approach on two surgical tasks: captioning and scene graph generation. The results prove that our gradient-based localized feature extraction methods effectively substitute the detector and feature extractor networks, allowing end-to-end model development with faster inference speed, essential for real-time surgical scene understanding tasks. The code is publicly available at https://github.com/PangWinnie0219/GradCAMDownstreamTask.**

***Index Terms*—Semantic Scene Understanding, Computer Vision for Medical Robotics, Medical Robots and Systems**

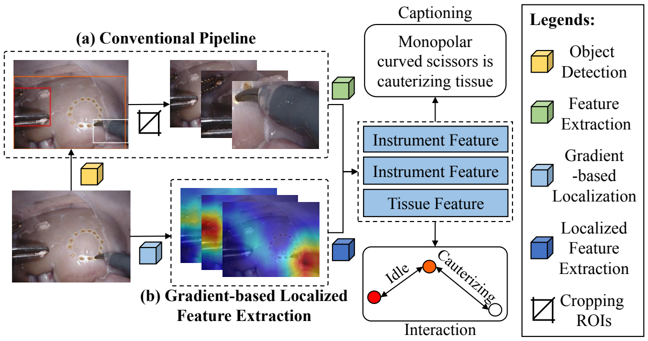

(a) Conventional Pipeline
Captioning
Monopolar curved scissors is cauterizing tissue
Instrument Feature
Instrument Feature
Tissue Feature
Idle
Cauterizing
Interaction
(b) Gradient-based Localized Feature Extraction
Legends:
Object Detection
Feature Extraction
Gradient -based Localization
Localized Feature Extraction
Cropping ROIs

Fig. 1. An overview of our proposed gradient-based localized feature extraction method (b), against the conventional approach of deep learning pipelines (a) for downstream tasks.

I. Introduction

Deep neural networks have increasingly become the standard tool leveraged for various problems in computer vision, ranging from object detection [1], [2], image classification [3], [4], semantic segmentation [5], natural language processing [6], [7] and scene graph generation [8], [9]. Despite its widespread success in these areas, the observed impact of its real-time application in robotic surgery is relatively insignificant. Surgical scene understanding which has shown great promise in robot-assisted minimally invasive surgery is associated with several downstream tasks such as scene graph generation for tool-tissue interaction detection [10] and surgical scene captioning or report generation [11]. However, most of such downstream task frameworks depend heavy on object detection and feature extraction (FE) pipelines for localized feature extraction to generate captions [11]–[14] or scene graphs [10], [15], [16]. Traditionally, the pipeline contains an object detector such as Faster-RCNN [2] to detect the object bounding box and a feature extraction backbone such as ResNet-50 [4] that extracts features from the detected object regions (cropped) for training the downstream task [12], [15] (as in Fig. 1(a)). This framework inherits limits such as: (i) requires massive amounts of manual bounding box annotations in addition to the captioning or scene graph labels to train the object detection pipeline [17], [18]; (ii) incorrect detection could severely affect the downstream tasks

Manuscript received: July, 6, 2022; Revised August, 31, 2022; Accepted November, 5, 2022.

This paper was recommended for publication by Editor Pietro Valdastri upon evaluation of the Associate Editor and Reviewers' comments.

This work was supported by the National Key R&D Program of China under Grant 2018YFB1307700 (with subprogram 2018YFB1307703) from the Ministry of Science and Technology (MOST) of China, Hong Kong Research Grants Council (RGC) Collaborative Research Fund (CRF C4026-21GF and C4063-18G), the Shun Hing Institute of Advanced Engineering (SHIAE project BME-p1-21, 8115064) at the Chinese University of Hong Kong (CUHK), Shenzhen-Hong Kong-Macau Technology Research Programme (Type C) Grant 202108233000303, and Singapore Academic Research Fund under Grant R397000353114.

[†] Winnie Pang and Mobarakol Islam are co-first authors.

[*] Corresponding author.

[1] Winnie Pang, Lalithkumar Seenivasan, Mengya Xu and Hongliang Ren are with Biomedical Engineering, National University of Singapore, Singapore. (winnie_pang@u.nus.edu,lalithkumar_s@u.nus.edu,mengya@u.nus.edu, hlren@ieee.org)

[2] Mobarakol Islam is with BioMedIA Group, Imperial College London, UK. (m.islam20@imperial.ac.uk)

[3] Sai Mitheran is with Dept. of ECE, National Institute of Technology, Tiruchirappalli, India. (108118084@nitt.edu)

[4] Hongliang Ren is with Electronic Engineering, The Chinese University of Hong Kong. (hlren@ee.cuhk.edu.hk)

Digital Object Identifier (DOI): see top of this page.

during both the training and inference stage, affecting their performance; (iii) the training is not end-to-end, leading to sub-optimal model convergence and performance; (iv) requires high computational resources limiting real-time application.

Grad-CAM (Gradient-weighted Class Activation Mapping) [19] is a widely used visual interpretability technique for trained neural networks. On a high level, it creates heatmaps to visualize regions in the input data that play a vital role in the model prediction and are often used to gain insights into why a model succeeds or fails in discriminative tasks. Exploiting its impressive localization ability, we utilize Grad-CAM and Grad-CAM++ [20] for gradient-based localized feature extraction to effectively replace object detection and feature extraction (FE) pipelines and allow end-to-end model training and inference for surgical downstream tasks with minimal computational load (Fig. 1(b)). We design three variants of localized feature extraction technique: (a) Localization and Naive FE: Employ gradient-based localization on a classification model for object region proposal and utilize ResNet50 to extract features from the cropped object regions. (b) Localization-aided FE: Employ gradient-based localization on a classification model for object region proposal and extract features from different layers of the same classification model. (c) Single-pass Localization-aided FE: Single-pass gradient-based localization and feature extraction from the classification model. Our key contributions and findings in this work are:

- We proposed a detector-free gradient-based localized feature extraction technique that allows end-to-end model training and inference for downstream tasks such as captioning and scene graph generation.
- We present a simple gradient-based localized feature extraction approach that effectively replaces object detection and feature extraction (FE) pipelines, removing the need for bounding box annotation, and allowing end-to-end downstream task training with significantly lower model parameters and inference in real-time.
- Through extensive experiments, we demonstrate that (i) The downstream tasks can be performed in an end-to-end manner, by replacing computationally expensive detection and feature extraction pipelines with only class-wise gradient localization; (ii) Our approach allows better localized-feature extraction and improves the performance of downstream tasks; (iii) The proposed technique surpasses the conventional approaches in both computational time and model parameters size.

II. Related-Work

A. Gradient-based Localization

Numerous works have been done in the direction of visual interpretability in localizing the important regions of an image. Class activation mapping (CAM) technique [21] is proposed to perform localization using Global Average Pooling (GAP) [22] in convolutional neural networks (CNNs). Later, Grad-CAM [19] is introduced to address the drawback of CAM that is only compatible with the classification CNN architectures. It uses gradients as weights to discriminate the regions in the image that contribute to the class prediction, and does not require modification in the network architecture. Further extensions of Grad-CAM include Grad-CAM++ [20], which adopts higher-order derivatives to improve the localization performance on small areas and multiple occurrences of the same object class in an image.

The applications of Grad-CAM in surgical robotics are mainly for the result visualization as part of explainability. Zhang et al. [23], Namazi et al. [24], and Jalal et al. [25] utilize Grad-CAM to visualize the significant regions focused by their models. However, relatively little attention has been paid to utilizing the localization ability of Grad-CAM in extracting the class-specific region features for downstream tasks such as image captioning and tool-tissue interaction detection.

B. Image Captioning

Captioning of surgical images in minimally invasive surgery allows automatic surgical report generation. This drastically reduces the burden on surgeons for documentation of operation procedures. Cornia et al. [13] introduces the M2T, the Transformer based memory-augmented encoder and decoder which has meshed connection with the encoder output. The model is improved by Xu et al. [11] with the introduction of the 1-dimensional Gaussian smoothing as curriculum learning into the encoder of M2T architecture and label smoothing loss. Although captioning model in [11] achieves outstanding performance in generating the surgical reports generation robotic surgery, the bounding box is still necessary for feature extraction (FE) before the captioning model. Recent works have attempted to study end-to-end captioning models using transformer-based architectures [26]–[28] where image patches are encoded into embeddings as grid representation, our approach utilizes gradient-based localized feature for the end-to-end captioning model.

C. Scene Graph Generation

The surgical scene graph generation is significant in real-time and post-surgical analysis, surgical skill assessment, as well as augmented tactile feedback. Works on generating scene graphs for scene understanding have increased recently. Human-object interaction detection task was theorised in a non-euclidean space and graph networks to construct scene graphs [9], [29]. Inspired by these works, Islam et al. [10] and Seenivasan et al. [16] improved and extended scene graphs to the medical domain for tool-tissue interaction detection in robotic surgery. While these models effectively detect tool-tissue interactions, they inherently rely on detection models to first predict bounding boxes for tissues and tools. This further makes these models not end-to-end training. Most recently, Li et al. [30] attempted to develop an end-to-end scene graph generation model. However, it is still a training detection head that inclines to introduce additional errors during graph interaction learning.

III. Proposed Method

A. Preliminaries

Grad-CAM [19] produces a class-specific heatmap for an image based on a weighted combination of feature maps

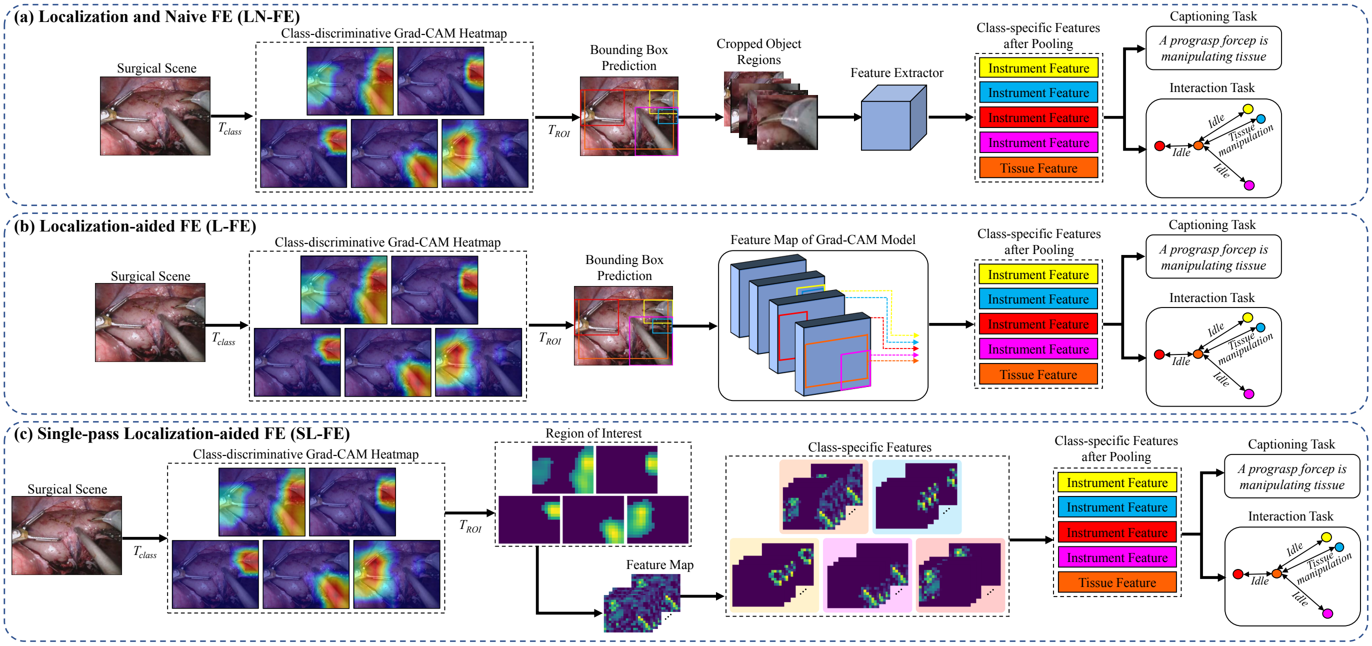


Fig. 2. The architecture of the proposed technique. A ResNet-50 [4] model is employed in our gradient-based localized feature extraction technique. We design three variants of localized feature extraction (FE) technique: (a) Localization and Naive FE (LN-FE), (b) Localization-aided FE (L-FE), and (c) Single-pass Localization-aided FE (SL-FE). These features can be used for downstream tasks such as scene graph generation and scene captioning.

from a chosen convolutional layer of the network. Moving beyond classical CAM [21], where the weights are chosen by pooling values from the last fully-connected layer of the network, Grad-CAM exploits the pooled gradients of class-specific logits y^c, w.r.t the chosen feature maps A^k_{ij} of a selected layer l, where $A^k \in \mathbb{R}^{h \times w}$. In this work, we select the *penultimate layer* of the trained ResNet-50 [4], with $k = 2048$. For every class c in the image, the heatmap H^c is generated using the following equation:

$$H^c = \text{ReLU}\left(\sum_k (\overbrace{\frac{1}{h \times w} \sum_i \sum_j \underbrace{\frac{\partial y^c}{\partial A^k_{ij}}}_{\text{gradients}}}^{\text{GAP}}) A^k \right) \quad (1)$$

where, ReLU refers to the Rectified Linear Unit activation function. In short, the Grad-CAM heatmap is yielded from the sum of the weighted feature map with ReLU activation, the weights are obtained from the gradients after global-average pooling. In this way, as the discriminative regions that influence class prediction are localized through the gradient weights of the feature map, the objects can be localized without the need to train an object detection model with bounding box annotations.

B. End-to-End Downstream Tasks with Gradient-based Localization

1) Gradient-based Localized Feature Extraction: A ResNet-50 [4] model is employed in our proposed technique for localized feature extraction. We design three different approaches for feature extraction (FE) from an input image. The extracted features are then utilized by the downstream task models for captioning or scene graph generation.

Localization and Naive FE (LN-FE): In this approach, we first generate the Grad-CAM heatmap of the same spatial size as the input and then produce bounding boxes from this heatmap using contour detection from the OpenCV Library [31]. As some false positive classes are observed in the heatmap predictions, we generate the Grad-CAM heatmap based on the predicted probability instead of the correctly predicted classes. The Grad-CAM heatmap for a class is obtained if its prediction probability exceeds a defined threshold, T_{class}. A threshold T_{ROI} is also applied to the heatmap to define the boundary for the region of interest (ROI). Finally, we crop the ROIs based on the detected contours, resize them and extract features using a feature extractor (Fig. 2(a)). While this approach is similar to the conventional approaches [12], [15], it eliminates the need for a detection pipeline.

Localization-aided FE (L-FE): This approach eliminates the need for cropping ROIs from the raw image and additional forward passes for feature extraction (FE), moving away from the conventional approach (Fig. 2(b)). Here, we extract the ROI features directly from the feature maps of the ResNet-50 model depending on the size of the bounding box $\triangle_B$, relative to the input image size $\triangle_I$. Intuitively, the features are extracted from initial layers of the ResNet-50 model for small bounding boxes, to extract a semantically-rich spatial representation of the region image. The feature map is chosen based on the equation:

$$F = \begin{cases} F_{LP}, & \triangle_B > \frac{1}{2}\triangle_I, \\ F_{L2}, & \frac{1}{4}\triangle_I < \triangle_B < \frac{1}{2}\triangle_I, \\ F_{L3}, & \triangle_B < \frac{1}{4}\triangle_I. \end{cases} \quad (2)$$

where, F_{LP}, F_{L2}, F_{L3} represent the $penultimate\ layer$, $second\,last\,layer$ and $third\,last\,layer$ of the feature maps.

Single-pass Localization-aided FE (SL-FE): To further reduce computation time, we propose a new approach to extract features based on the class-discriminative Grad-CAM heatmap, eliminating the need for the OpenCV library. Similar to the bounding box generation, a threshold T_{ROI} is applied to the class-specific Grad-CAM heatmap H^c to highlight only the ROIs (Fig. 2(c)).

$$M_{H^c} = \left[m_{i,j} = \begin{cases} 1, & \text{if} \quad H^c_{i,j} > T_{ROI} \\ 0, & \text{else} \end{cases} \right]_{(i,j) \in H^c} \tag{3}$$

where, M_{H^c} refers to the heatmap obtained after thresholding. M_{H^c} is expanded to the length of channel dimension k of the feature maps A. It is then masked with the feature maps, to extract class-specific features F^c.

$$F^c = [(M_{H^c})_{\times k}] \odot A \tag{4}$$

Adaptive average pooling is adopted to transform all feature vectors F^c of arbitrary size to a fixed desired size (512), which are then used by the downstream task models.

2) Scene Captioning: Utilizing the gradient-based localized features, the surgical scene captioning is achieved using Mesh-Memory Transformer (M2T) [11], [13], a transformer-based multi-layer encoder-decoder model. The encoder comprises stacks of self-attention layers augmented with memory and feed-forward layers to learn the relationship between regions of inputs. The decoder consists of cross-attention for encoder outputs and self-attention on words to learn the multi-level representation from the input and generate the output sequences. The output sequences encode the probability of words in the dictionary.

3) Scene Graph Generation: The SG-Transformer [8] scene graph generation model is modified to perform tool-tissue interaction detection. A transformer architecture is used for both the object and relation encoder of the SG-Transformer. Object embeddings are obtained from the object encoder with the visual features. On the other hand, the pairwise object features, word embedding [32] of the predicted object labels, as well as the object embedding are fed as inputs to the relation encoder. The relation encoder output is then concatenated with the object embedding of the subject-object pair and is decoded by a fully connected layer with softmax activation.

IV. Experiments

A. Dataset

1) EndoVis18: Robotic instrument segmentation dataset from MICCAI endoscopic vision challenge 2018 [33] (EndoVis18) contains 15 robotic surgical videos with 8 instruments present in these surgical procedures. The dataset is originally annotated for segmentation, and further tool-tissue interaction and captioning annotations are provided in [16] and [11], respectively. There are 13 different interaction classes and corresponding captions in the dataset. We utilize 3 video sequences (1, 5, 16) for testing and the remaining sequences for training by following previous works [11], [16]. We generate multi-label tissue and instrument classification labels from the segmentation annotation. In total, there are 9 classes, and a frame can contain 4 classes at a time.

2) Cholec80: Cholec80 surgical workflow dataset [34] (Cholec80) includes 80 videos with 7 instruments for robotic surgery. The original dataset provides surgical workflow (phase) and multi-label tool presence labels. An extra *tissue* label is added to all video frames for tissue localization by the Grad-CAM model for the downstream tasks. We split the first 40 videos following an 80/20 ratio for training and testing. The videos are carefully chosen to allow balance class in the validation set (Video 05, 11, 12, 17, 19, 26, 27, and 31). In our experiments, we adopt only the tool presence labels.

B. Implementation Details

The ResNet-50 [4] model uses training splits from the EndoVis18 and Cholec80 datasets and is evaluated only on the EndoVis18 dataset. It is trained for multi-label surgical tool classification task. The model is loaded with ImageNet [35] pre-trained weights, and is trained for 200 epochs using the Stochastic Gradient Descent (SGD) optimizer with weight decay of 0.0001, and momentum of 0.9. The learning rate is 0.001 and is reduced by 0.95 with patience of 3 epochs. Multi-Label Soft Margin Loss is adopted to compute the multi-label classification loss. Input images from both datasets are resized to 256 x 320 by preserving the aspect ratio. All networks are implemented in PyTorch and trained using NVIDIA RTX 2080Ti GPUs. The captioning and interaction model are implemented following the works from [13][1] and [8][2].

V. Results

A. Performance Analysis

The performance of our proposed approach and its contribution to enhancing the performance of downstream tasks is analysed quantitatively, qualitatively and computationally. Firstly, the ResNet-50 model trained for instrument classification achieves a classification mean average precision (mAP) of 0.6470. Secondly, the localization (object detection) performance of our approach that utilizes the trained ResNet-50 (instrument classification) is evaluated using the detection mAP at a threshold of 0.5 (mAP@0.5) against SOTA models (Faster RCNN [2], YOLOv5 [36]) and YOLOv7 [37]. Thirdly, the performance of downstream tasks (scene captioning and scene graph generation) trained on features extracted using our approach is benchmarked against traditional pipelines (SOTA object detection model + ResNet-50 feature extraction (FE) network). The captioning performance is evaluated using BLEU-1, BLEU-4 [38] and CIDEr [39]. The scene graph generation performance is evaluated using Recall and balanced mean Recall (mRecall) [40], [41] metrics.

Detection: As one of our proposed variants (SL-FE) does not generate bounding boxes, its performance is not included for evaluation on object detection. From TABLE I, it is

[1] https://github.com/aimagelab/meshed-memory-transformer

[2] https://github.com/CYVincent/Scene-Graph-Transformer-CogTree

This article has been accepted for publication in IEEE Robotics and Automation Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LRA.2022.3221310

TABLE I
PERFORMANCE COMPARISON OF OUR GRADIENT-BASED LOCALIZED FEATURE EXTRACTION METHOD AND ITS VARIANTS AGAINST THE CONVENTIONAL OBJECT DETECTION (DET.) APPROACHES IN BOUNDING BOX DETECTION, SCENE CAPTIONING [11], AND TOOL-TISSUE INTERACTION DETECTION [8]. FE, RN50, MRECALL REFER TO FEATURE EXTRACTOR, RESNET-50 [4] AND MEANRECALL.

Method	**Model**		**FPS**	**No. of Params**	**Detection**	**Captioning** [11]			**Interaction** [8]	
	Det. Model	**FE**			**mAP@0.5**	**BLEU-1**	**BLEU-4**	**CIDEr**	**Recall**	**mRecall**
Conventional approach	Faster RCNN [2]	RN50	18.88	41.12 M	0.5538	0.4822	0.3182	2.04	0.2815	0.1530
	YOLOv5x [36]		25.05	109.73 M	0.5640	0.5767	0.4122	2.90	0.2940	0.1623
	YOLOv5s [36]		34.15	30.54 M	0.5450	0.5932	0.4443	2.79	0.2822	0.1413
	YOLOv7-E6E [37]		19.89	188.49 M	0.5670	0.6180	0.4566	2.85	0.3412	0.1386
	YOLOv7 [37]		35.78	60.03 M	0.5290	0.6129	0.4470	2.84	0.3055	0.1555
Ours (LN-FE)	Grad-CAM [19]	RN50	36.26	47.04 M	**0.6396**	0.6243	0.4587	3.41	0.2850	**0.1885**
Ours (L-FE)		✗	65.79	**23.51 M**		0.5984	0.4425	2.87	0.3272	0.1794
Ours (SL-FE)			**67.59**		✗	0.5900	0.4368	3.22	0.3523	0.1716
Ours (LN-FE)	Grad-CAM++ [20]	RN50	23.58	47.04 M	0.6237	**0.6532**	**0.5059**	**4.02**	0.3322	0.1444
Ours (L-FE)		✗	33.55	**23.51 M**		0.5731	0.4096	2.94	0.3358	0.1623
Ours (SL-FE)			35.10		✗	0.5530	0.3859	2.84	**0.3529**	0.1717

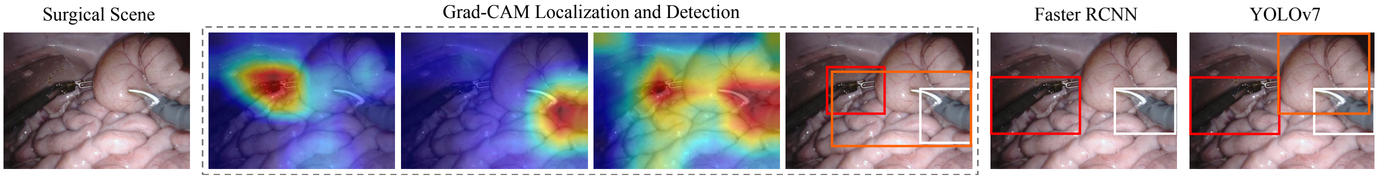
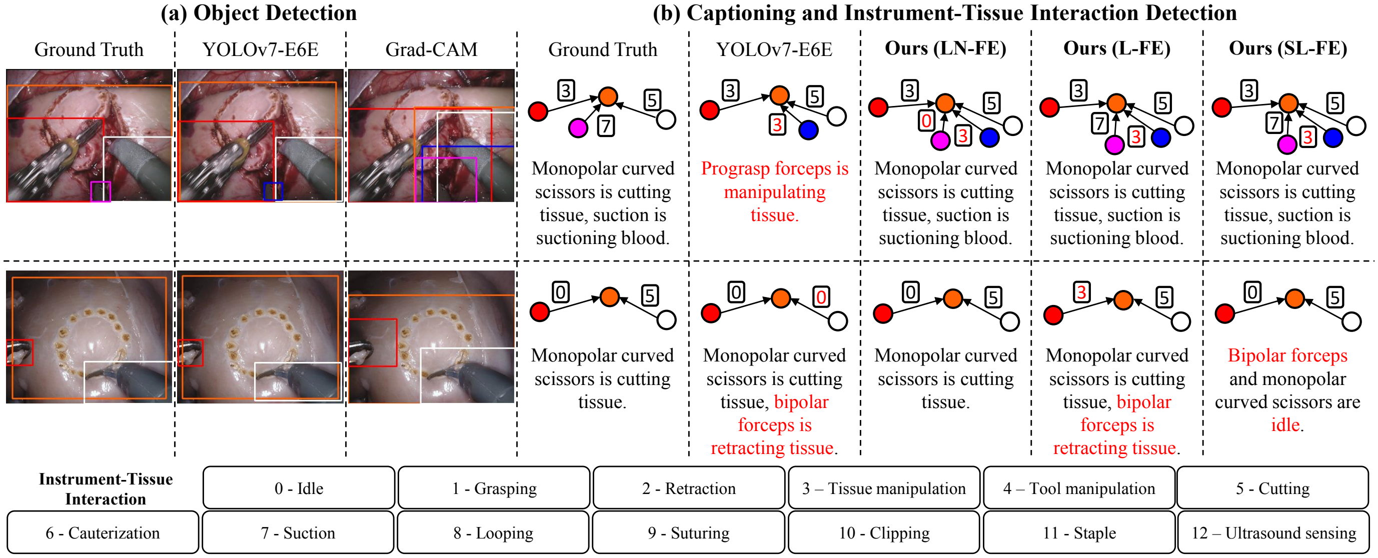


Fig. 3. Visualization of the gradient-based localization and detection. Object detection performance is also compared with SOTA models (Faster RCNN and YOLOv7). First image shows the original surgical image with a bipolar forceps (left) and monopolar curved scissors (right). Second, third and forth images show the Grad-CAM heatmaps of the 'bipolar forceps', 'monopolar curved scissors' and 'kidney' respectively. Forth figure shows the bounding boxes generated from the Grad-CAM heatmaps. The last two figures show the predicted bounding boxes from Faster RCNN and YOLOv7 models respectively.



Ground Truth: Monopolar curved scissors is cutting tissue, suction is suctioning blood.
YOLOv7-E6E: Prograsp forceps is manipulating tissue.
Ours (LN-FE): Monopolar curved scissors is cutting tissue, suction is suctioning blood.
Ours (L-FE): Monopolar curved scissors is cutting tissue, suction is suctioning blood.
Ours (SL-FE): Monopolar curved scissors is cutting tissue, suction is suctioning blood.

Ground Truth: Monopolar curved scissors is cutting tissue.
YOLOv7-E6E: Monopolar curved scissors is cutting tissue, bipolar forceps is retracting tissue.
Ours (LN-FE): Monopolar curved scissors is cutting tissue.
Ours (L-FE): Monopolar curved scissors is cutting tissue, bipolar forceps is retracting tissue.
Ours (SL-FE): Bipolar forceps and monopolar curved scissors are idle.

Fig. 4. Visualization of (a) Object detection with the conventional approach (YOLOv7-E6E) and the gradient-based localization (Grad-CAM). (b) Comparison of the scene captioning and interaction detection performance of our proposed feature extraction variants and the conventional object detection approach (YOLOv7-E6E). The output of the captioning model is listed at the bottom of the interaction detection results on EndoVis18 dataset. The legend at the bottom provides tool-tissue interaction labels. The indices and words marked in red denote incorrect predictions. Additional qualitative evaluations can be found in the supplementary video.

observed that our gradient-based localization approaches significantly surpass SOTA object detection models in terms of mAP@0.5, with fewer model parameters and higher FPS. It is worth noting that the ResNet-50 model utilized in the gradient-based localized approach is trained solely on the instrument labels while the conventional approach is trained with bounding box annotations, demonstrating the ability to generate bounding boxes without expensive human annotations. An example of the gradient-based localization vs SOTA detection model is shown in Fig. 3 and Fig. 4(a). The comparative

TABLE II
Improvement of our proposed approaches in captioning tasks by comparing existing captioning model (M2T [13]) and surgical captioning model (Xu et al. [11]). For a fair comparison, we use predicted bounding box instead of groundtruth bounding box for feature extraction and avoid the incremental learning module to train Xu et al. [11].

Detection Model	Captioning Model	BLEU-1	BLEU-4	CIDEr
Faster RCNN [2]	M2T [13]	0.5250	0.3679	2.24
	Xu et al. [11]	0.4822	0.3182	2.04
YOLOv5x [36]	M2T [13]	0.5435	0.3942	2.88
	Xu et al. [11]	0.5767	0.4122	2.90
YOLOv5s [36]	M2T [13]	0.5836	0.4279	2.72
	Xu et al. [11]	0.5932	0.4443	2.79
YOLOv7-E6E [37]	M2T [13]	0.6145	0.4665	3.09
	Xu et al. [11]	0.6180	0.4566	2.85
YOLOv7 [37]	M2T [13]	0.6167	0.4573	2.77
	Xu et al. [11]	0.6129	0.4470	2.84
Ours (LN-FE)	M2T [13]	0.6432	0.5047	3.94
	Xu et al. [11]	**0.6532**	**0.5059**	**4.02**
Ours (L-FE)	M2T [13]	0.5449	0.3663	2.50
	Xu et al. [11]	0.5731	0.4096	2.94
Ours (SL-FE)	M2T [13]	0.5116	0.4232	2.70
	Xu et al. [11]	0.5530	0.3859	2.84

figures of Grad-CAM and Grad-CAM++ localizations can be found in the supplementary video.

Scene Captioning: In the downstream scene captioning task, the models trained from features extracted using our approach performed on par with models training from features extracted using traditional SOTA pipelines while requiring lesser model parameters and attaining higher FPS (TABLE I). It is worth noting that the model trained using our approach with Grad-CAM++ localization achieved the best performance on BLEU-1, BLEU-4 and CIDEr. This implies that our gradient-based localized features extraction method can provide optimal outcomes in downstream tasks. Similar performance was observed qualitatively in Fig. 4(b). Furthermore, the effects of our proposed approaches are investigated using different detection methods and existing captioning models such as M2T [13], and Xu et al. [11] (TABLE II). The results suggest the significant performance gain with the proposed methods, specifically with the variant of LN-FE. Although YOLOv7 [37] yields competitive prediction, it requires much more computational resources.

Scene Graph Generation: Similar to the scene captioning task, in the downstream scene graph generation task, the models trained on features extracted using our approaches outperformed or performed on par with models training using traditional SOTA pipelines. Models trained with our SL-FE approach with both Grad-CAM and Grad-CAM++ localizations achieved the highest recall. mRecall calculates the recall of each predicted class and obtains the mean of all the predicted classes, evaluating if the model is biased towards the dominant class. High mRecall is also observed for models trained on features extracted using our SL-FE approach. As our SL-FE approach doesn't provide bounding boxes and allows the features to be extracted only from the highlighted class-discriminative regions instead of the region from the whole bounding box, our approach could have helped filter noises (irrelevant features) and improve the model's performance and generalization in tool-tissue interaction detection.

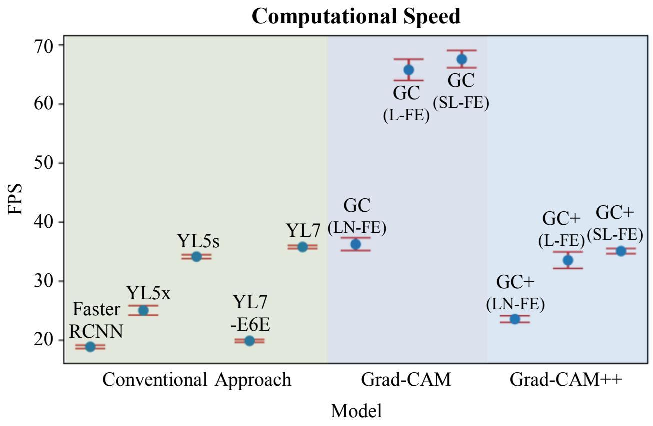


Fig. 5. Computational efficiency (in terms of FPS) of the SOTA approaches and our proposed feature extraction methods with Grad-CAM and Grad-CAM++ localizations. YL, GC, GC+ refer to YOLO, Grad-CAM and Grad-CAM++

B. Computational Analysis

Under the same hardware environment, the computational efficiency (in terms of FPS) of our proposed localized feature extraction approach with Grad-CAM and Grad-CAM++ localizations against other SOTA approaches is also studied (Fig. 5). Despite having a similar feature extraction (FE) setup, our LN-FE approach with a gradient-based localization technique extracts ROI features faster than the two-stage detector (Faster RCNN) as well as the single-stage detector of a larger scale (YOLOv5x and YOLOv7-E6E) and obtains comparable speed with the single-stage detector of smaller scale (YOLOv5s and YOLOv7). Approaches using Grad-CAM++ localization has relatively lower computational speed compared to those using Grad-CAM localization due to the additional computation on the second-order gradients. Significant improvements in computation speed can be observed in both L-FE approach and SL-FE approach as the additional forward pass of ROIs for feature extraction is not required. By removing the need for bounding box generation and thus eliminating the utilization of the OpenCV Library, the SL-FE approach achieves a further efficiency boost by 1.5 FPS. This method provides promising inference speed for models with our localized feature extraction approach to be deployed for real-time surgical applications.

C. Ablation Study

Ablation studies on the effects of different training datasets and parameters on our proposed Single-pass Localization-aided FE (SL-FE) approach with Grad-CAM localization are studied.

TABLE III

VARIATIONS IN THE TRAINING DATASETS OF THE CLASSIFICATION MODEL (RESNET-50). TO ACCOUNT FOR ALL INSTRUMENTS IN BOTH DATASETS, WE ADJUST THE TOOL PRESENCE ANNOTATIONS OF THE ENDOVIS18 DATASET BY EXPANDING FROM 9 CLASSES TO 11 CLASSES. CLASS LABELS MARKED WITH * ARE ONLY PRESENT IN THE CHOLEC80 DATASET. T: TRAINING ON THE RESNET-50 INITIALIZED WITH IMAGENET PRETRAINED WEIGHTS. F: FINE-TUNING FROM THE RESNET-50 PRE-TRAINED ON CHOLEC80.

Model	Cholec80	EndoVis18	Class Label
GC-A	✗	T	bipolar forceps, prograsp forceps, large needle driver, clip applier, monopolar curved scissors, suction, ultrasound probe, stapler, tissue
GC-B	T	F	
GC-C			bipolar forceps, prograsp forceps, large needle driver, clip applier, monopolar curved scissors, suction, ultrasound probe, stapler, hook*, specimen bag*, tissue
GC-D	T		

TABLE IV

CAPTIONING AND INTERACTION DETECTION PERFORMANCE FOR OUR SL-FE APPROACH WITH DIFFERENT THRESHOLD T_{ROI}.

T_{ROI}	Captioning		Interaction	
	BLEU-4	CIDEr	Recall	mRecall
0.1	0.4368	**3.22**	0.2841	**0.2034**
0.3	**0.4792**	3.18	**0.3175**	0.1467
0.5	0.4339	2.70	0.2919	0.1409

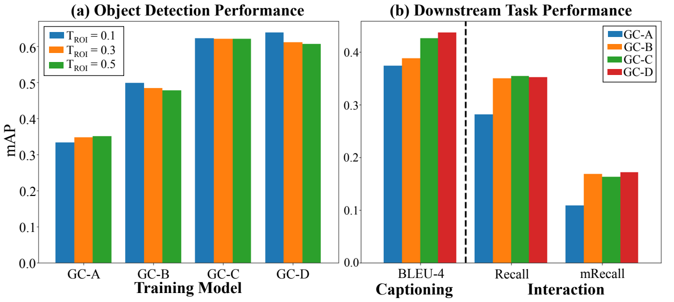


Fig. 6. (a) Object detection performance of the ResNet-50 model with gradient-based localization in four variations of training datasets, under different heatmap threshold T_{ROI}. (b) Captioning and interaction detection performance with the localized features extraction from the ResNet-50 model with different variations of training datasets.

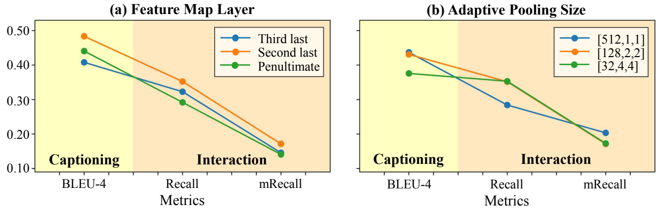


Fig. 7. Captioning and interaction detection performance for our localization (local.) method with features extracted from different layers of the ResNet-50 model, and various adaptive average pooling sizes.

1) Training datasets: The ImageNet pre-trained ResNet-50 model is trained with four variations (TABLE III) of the EndoVis18 and Cholec80 datasets and evaluated on the EndoVis18 dataset. Among all, *GC-D* achieves the best overall results in both detection and downstream task performance (Fig. 6). This result highlights the possibility of boosting downstream task performance by training the model together with other similar datasets: even in the absence of bounding box annotations.

2) Grad-CAM Heatmap Threshold: A threshold T_{ROI} of 0.1, 0.3, and 0.5 is selected to study the effect of the threshold applied to the Grad-CAM heatmap in extracting useful features for downstream task performances. As shown in TABLE IV, the performance on both downstream tasks drops when T_{ROI} = 0.5. This may arise due to the reduced ROI area, thus resulting in loss of features being extracted. Ultimately, proper selection of the threshold T_{ROI} may boost model performance, but the overall performance on downstream tasks remains stable with varying thresholds.

3) Feature Map Layer: We study the downstream task performance with features extracted from *penultimate layer*, *second last layer*, and *third last layer* of the ResNet-50 model, as shown in Fig. 7(a). For both captioning and graph-based interaction detection, features extracted from *second last layer* perform better than the features extracted from the other layers, as more spatial semantics are retained.

4) Adaptive Pooling Size: We also investigate the impact of adaptive pooling sizes on the extracted features for downstream task performance. After adaptive average pooling, the features are flattened (size = 512) for further processing. The pooling size of [512,1,1] shows improved task performance compared to the pooling sizes of [128,2,2] and [32,4,4] in BLEU-4 of captioning task and meanRecall in interaction task as illustrated in Fig. 7(b).

VI. DISCUSSION AND CONCLUSION

In this paper, we leverage the localization ability of Grad-CAM to establish gradient-based localized feature extraction (FE) approaches. We introduce three variants to extract localized features using a ResNet-50 model, eliminating the need for the computationally heavy object detector and feature extractor pipeline, and enabling end-to-end model development for surgical downstream tasks. Additionally, our Single-pass Localization-aided FE (SL-FE) approach allows region features to be extracted directly from the classification model based on its gradient weights, eliminating the need for bounding box generation for feature extraction. Moreover, the training procedures with our feature extraction approaches require only the instrument and tissue labels. This serves as a significant advantage for ROI-centric downstream tasks such as scene captioning and scene graph generation tasks as bounding box annotations for medical images are exceptionally costly to procure, given the time, human resources, and limited expertise available from professionals. Furthermore, our approach requires significantly less computational power and allows the development of real-time applications. We demonstrate the robustness of our proposed framework through several experiments on detection and surgical downstream tasks in Section V. We also explore extending our proposed approaches with Grad-CAM++ [20], improving object localization for images with multiple occurrences of the same class. Future work includes integrating temporal information from videos and improving the model robustness towards various surgical video datasets.

REFERENCES

[1] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[6] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[7] I. V. Tetko, P. Karpov, R. Van Deursen, and G. Godin, "State-of-the-art augmented nlp transformer models for direct and single-step retrosynthesis," *Nature communications*, vol. 11, no. 1, pp. 1–11, 2020.

[8] J. Yu, Y. Chai, Y. Wang, Y. Hu, and Q. Wu, "Cogtree: Cognition tree loss for unbiased scene graph generation," *arXiv preprint arXiv:2009.07526*, 2020.

[9] Z. Liang, J. Rojas, J. Liu, and Y. Guan, "Visual-semantic graph attention networks for human-object interaction detection," *arXiv preprint arXiv:2001.02302*, 2020.

[10] M. Islam, L. Seenivasan, L. C. Ming, and H. Ren, "Learning and reasoning with the graph structure representation in robotic surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 627–636.

[11] M. Xu, M. Islam, C. M. Lim, and H. Ren, "Class-incremental domain adaptation with smoothing and calibration for surgical report generation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 269–278.

[12] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[13] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 578–10 587.

[14] M. Cornia, L. Baraldi, and R. Cucchiara, "Explaining transformer-based image captioning models: An empirical analysis," *AI Communications*, no. Preprint, pp. 1–19, 2021.

[15] Z. Liang, J. Rojas, J. Liu, and Y. Guan, "Visual-semantic graph attention networks for human-object interaction detection," *arXiv preprint arXiv:2001.02302*, 2020.

[16] L. Seenivasan, S. Mitheran, M. Islam, and H. Ren, "Global-reasoned multi-task learning model for surgical scene understanding," *IEEE Robotics and Automation Letters*, 2022.

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[18] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017.

[19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[20] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. Balasubramanian, "Grad-cam++: Improved visual explanations for deep convolutional networks. arxiv 2017," *arXiv preprint arXiv:1710.11063*.

[21] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929, 2016.

[22] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[23] D. Zhang, R. Wang, and B. Lo, "Surgical gesture recognition based on bidirectional multi-layer independently rnn with explainable spatial feature extraction," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 1350–1356.

[24] B. Namazi, G. Sankaranarayanan, and V. Devarajan, "Laptool-net: a contextual detector of surgical tools in laparoscopic videos based on recurrent convolutional neural networks," *arXiv preprint arXiv:1905.08983*, 2019.

[25] N. A. Jalal, H. Arabian, T. A. Alshirbaji, P. D. Docherty, T. Neumuth, and K. Moeller, "Analysing attention convolutional neural network for surgical tool localisation: a feasibility study," *Current Directions in Biomedical Engineering*, vol. 8, no. 2, pp. 548–551, 2022.

[26] Z. Fang, J. Wang, X. Hu, L. Liang, Z. Gan, L. Wang, Y. Yang, and Z. Liu, "Injecting semantic concepts into end-to-end image captioning," *ArXiv*, vol. abs/2112.05230, 2021.

[27] M. Xu, M. Islam, and H. Ren, "Rethinking surgical captioning: End-to-end window-based mlp transformer using patches," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 376–386.

[28] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8739–8748.

[29] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," *arXiv preprint arXiv:1808.07962*, 2018.

[30] R. Li, S. Zhang, and X. He, "Sgtr: End-to-end scene graph generation with transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 486–19 496.

[31] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[32] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[33] M. Allan, S. Kondo, S. Bodenstedt, S. Leger, R. Kadkhodamohammadi, I. Luengo, F. Fuentes, E. Flouty, A. Mohammed, M. Pedersen *et al.*, "2018 robotic scene segmentation challenge," *arXiv preprint arXiv:2001.11190*, 2020.

[34] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy, "Endonet: a deep architecture for recognition tasks on laparoscopic videos," *IEEE transactions on medical imaging*, vol. 36, no. 1, pp. 86–97, 2016.

[35] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[36] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, and Y. K. et al., "ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference," Feb. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6222936

[37] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *arXiv preprint arXiv:2207.02696*, 2022.

[38] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[39] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[40] T. Chen, W. Yu, R. Chen, and L. Lin, "Knowledge-embedded routing network for scene graph generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6163–6171.

[41] K. Tang, H. Zhang, B. Wu, W. Luo, and W. Liu, "Learning to compose dynamic tree structures for visual contexts," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6619–6628.